

Improving information filtering via network manipulation

FUGUO ZHANG^{1,2} AND AN ZENG^{2(a)}

¹ School of Information and Technology, Jiangxi University of Finance and Economics, Nanchang 330013, P.R. China

² Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland

PACS 89.75.-k – Complex systems

PACS 89.65.-s – Social and economic systems

PACS 89.20.Ff – Computer science and technology

Abstract. - Recommender system is a very promising way to address the problem of overabundant information for online users. Though the information filtering for the online commercial systems received much attention recently, almost all of the previous works are dedicated to design new algorithms and consider the user-item bipartite networks as given and constant information. However, many problems for recommender systems such as the cold-start problem (i.e. low recommendation accuracy for the small degree items) are actually due to the limitation of the underlying user-item bipartite networks. In this letter, we propose a strategy to enhance the performance of the already existing recommendation algorithms by directly manipulating the user-item bipartite networks, namely adding some virtual connections to the networks. Numerical analyses on two benchmark data sets, *MovieLens* and *Netflix*, show that our method can remarkably improve the recommendation performance. Specifically, it not only improve the recommendations accuracy (especially for the small degree items), but also help the recommender systems generate more diverse and novel recommendations.

Introduction. – In the Internet era, the rapid growth of the World-Wide-Web leads to a serious problem of information overload: people are now facing too many choices to be able to find out those most relevant ones [1]. So far, the most promising way to efficiently filter the abundant information is to employ the personalized recommendations [2, 3]. That is to say, using the personal history record of a user to uncover his preference and to return each user with the most relevant items according to his taste [4]. For instances, youtube.com uses people's video viewing record to provide individual suggestions for their potential interested videos.

There are already many recommendation algorithms for the online user-item commercial systems. Among these algorithms, the simplest one is the popularity-based recommendations, which recommend the most popular items to users. However, such recommendations are not personalized so that identical items are recommended to individuals with far different tastes. By comparison, the collaborative filtering makes use of collective data from

individual preferences to provide personalized recommendations [5, 6]. Recently, recommendation algorithms have been proposed from a physics perspective. For example, the process of mass diffusion (MD) was applied on the user-item bipartite networks to explore items of potential interest for a user [7]. The mass diffusion algorithm outperforms the previous ones in the recommendation accuracy. However, such method is still biased to popular items even if individual preferences are considered. An alternative approach, based on the heat conduction (HC) on the user-item graphs, was thus introduced [8]. This algorithm provides users with many novel items and leads to diverse recommendations among users. However, HC has low accuracy compared with MD. This drawback is eventually solved by combining MD with HC in a hybrid approach, which can be well-tuned to obtain significant improvement in both recommendation accuracy and item diversity [9]. More Recently, the long term influence of such hybrid approach on network evolution has been studied [10].

However, all these methods are focusing on improve the recommendation all from the system point of view. The

(a) an.zeng@unifr.ch

recommendation on the items with little information are actually still a critical challenging [4]. For the fresh or unpopular items (also called niche items), it is very difficult to predict the potential users who are going to interest in them due to lacking of historical record. Such problem is always referred as cold-start problem and many researches have been dedicated to solve this problem. Related works concerning this issue are mainly based on modifying the existing methods by introducing some parameters [11–13].

Different from previous works, we tried to solve the cold-start problem through a very fundamental way in this letter. Instead of designing a new recommendation algorithm, we make use of the MD algorithm and solve the cold-start problem by directly manipulating the underlying user-item bipartite networks [14, 15]. Actually, the idea of the network manipulation has been applied to enhance many kinds of network functions such as synchronization [16, 17], traffic dynamics [18], percolation [19, 20], navigation [21] and so on. In our case, we first analyze the historical record of each item and accordingly add some virtual connections to the networks (especially for the small degree items) to provide the recommendation algorithm with more information. By using the MD algorithm, we find that the recommendation accuracy for the small degree items can be largely enhanced after manipulating the networks. The further test on the overall recommendation metrics, our method are shown to help the MD algorithm outperform the hybrid approach of the MD and HC algorithms in both recommendation accuracy and diversity.

Recommendation algorithms. – Online commercial systems can be well described by the user-item bipartite networks. If a user collects an item, a link is drawn between them. Specifically, we consider a system of N users and M items represented by a bipartite network with adjacency matrix A , where the element $a_{i\alpha} = 1$ if a user i has collected an item α , and $a_{i\alpha} = 0$ otherwise (throughout this paper we use Greek and Latin letters, respectively, for item- and user-related indices).

There are many recommendation algorithms. In this letter, we mainly consider the Mass Diffusion (MD), Heat Conduction (HC) and the corresponding hybrid algorithms of these two algorithms (Hybrid). We first briefly describe these algorithms.

For a target user i , the MD algorithm [7] starts by assigning one unit of resource to each item collected by i , and redistributes the resource through the user-item network. We denote the vector \mathbf{f}^i as the initial resources on items, where the α -th component f_α^i is the resource possessed by item α . Recommendations for the user i are obtained by setting the elements in \mathbf{f}^i to be $f_\alpha^i = a_{i\alpha}$, in accordance with the items the user has already collected. The redistribution is represented by $\tilde{\mathbf{f}} = W\mathbf{f}^i$, where

$$W_{\alpha\beta} = \frac{1}{k_\beta} \sum_{j=1}^N \frac{a_{j\alpha}a_{j\beta}}{k_j}, \quad (1)$$

is the diffusion matrix, with $k_\beta = \sum_{l=1}^N a_{l\beta}$ and $k_j = \sum_{\gamma=1}^M a_{j\gamma}$ denoting the degree of item β and user j respectively. The resulting recommendation list of uncollected items is then sorted according to \tilde{f}_α^i in descending order. Physically, the diffusion is equivalent to a three-step random walk starting with k_i units of resources on the target user i . The *recommendation score* of an item is taken to be its amount of gathered resources after the diffusion. This algorithm was shown to enjoy a high recommendation accuracy.

The HC algorithm [8] works similar to the MD algorithm, but instead follows a conductive process represented by

$$W_{\alpha\beta} = \frac{1}{k_\alpha} \sum_{j=1}^N \frac{a_{j\alpha}a_{j\beta}}{k_j}. \quad (2)$$

Physically, the recommendation scores can be interpreted as the temperature of an item, which is the average temperature of its nearest neighborhood, i.e. its connected users. The higher the temperature of an item, the higher its recommendation score. By using this algorithm, the items with small degree can receive relatively high recommendation score and finally be promoted to appear in the top recommendation list.

The hybrid algorithm of MD and HC was proposed in [9], with the new recommendation score \tilde{h}_α given by

$$\tilde{h}_\alpha = \lambda \frac{\tilde{f}_\alpha^{\text{MD}}}{\text{Max}(\tilde{f}^{\text{MD}})} + (1 - \lambda) \frac{\tilde{f}_\alpha^{\text{HC}}}{\text{Max}(\tilde{f}^{\text{HC}})}. \quad (3)$$

where the parameter λ adjusts the relative weight between the two algorithms. When λ increases from 0 to 1, the hybrid algorithm changes gradually from HC to MD. Such hybrid approach was shown to achieve both accurate and diverse recommendation.

Data. – In order to test the performance of the recommendation results, we use two benchmark data sets in this letter. The first one is the MovieLens data [22] which has 1,682 movies (items) and 943 users. The other is Netflix data [23] consisting of 10,000 users and 6,000 movies. The data sets are random samplings of users activity records in these two online systems. In both data sets, users can vote movies by giving different rating levels from 1 to 5 (i.e. worst to best). Here, only the rating larger than 2 are considered as a link. After this preliminary filtering, there are finally 82,520 links in movielens data and 701,947 links in the netflix data. Each data is then randomly divided into two parts: the training set (E^T) and the probe set (E^P). The training set contains 90% of the original data and the recommendation algorithm runs on it. The probe set has the remaining 10% of the data and will be used to test the performance of the recommendation results.

The network manipulating method. – The network manipulating (NM) method takes place after dividing the data to E^P and E^T . The main idea of NM is to add some virtual links to the training set E^T , so that the

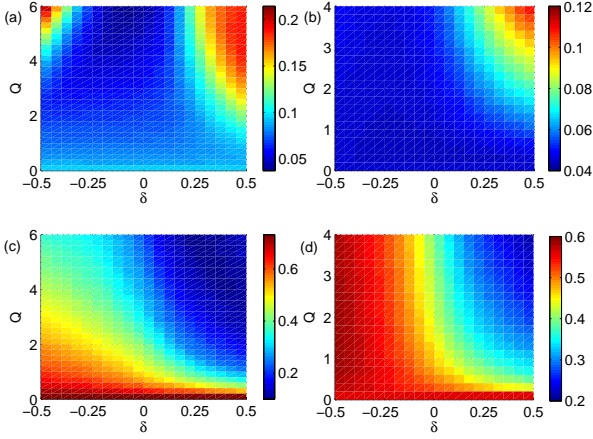


Fig. 1: (Color online) The overall ranking score R under different δ and Q in (a) Movielens and (b) Netflix, and the local ranking score $R_{k \leq 5}$ under different δ and Q in (c) Movielens and (d) Netflix.

niche items will have more information to provide to the recommendation algorithm. Denoting Q as the fraction of links added, the total number of virtual links will be $Q|E^T|$. For each item, the probability to receive virtual links is related to its degree, i.e. $p_\alpha \propto k_\alpha^{-\delta}$ where δ is a tunable parameter. when $\delta > 0$, the items with smaller degree tend to receive more links, and vice versa. Supposing an item α is selected to receive a link, the virtual link will connect to the user who enjoys the highest average similarity to the already existing selectors of the item α . In this letter, the similarity is calculated by *Salton Index* [24] as

$$s_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\sqrt{k_i k_j}}. \quad (4)$$

where $\Gamma(i)$ denotes the set of neighbors of user i . After adding virtual connections to the networks, we will employ the MD algorithm to do the recommendation and this combination is denoted as “MD with NM”. In the following discussion, we will compare this method to the original MD algorithm and the hybrid approach of MD and HC.

Metrics for recommendation. — An effective recommendation should be able to accurately find the items that users like. In order to measure the recommendation accuracy, we make use of *ranking score* (R). Specifically, R measures whether the ordering of the items in the recommendation list matches the users’ real preference. As discussed above, the recommender system will provide each user with a ranking list which contains all his uncollected items. For a target user i , we calculate the position for each of his link in the probe set. Supposing one of his uncollected item α is ranked at the 5th place and the total number of his uncollected items is 100, the ranking score $R_{i\alpha}$ will be 0.05. In a good recommendation, the items

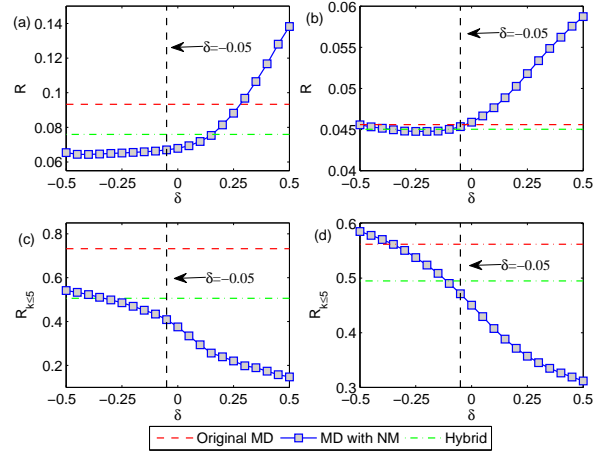


Fig. 2: (Color online) The overall ranking score R and the local ranking score $R_{k \leq 5}$ under different δ when $Q = 2$ in (a), (c) Movielens and $Q = 1$ in (b), (d) Netflix. The vertical dash line is the optimal δ we used.

in the probe set should be ranked higher, so that R will be smaller. Therefore, the mean value of the R over all the user-item relations in the probe set can be used to evaluate the recommendation accuracy as

$$R = \frac{1}{|E^P|} \sum_{i\alpha \in E^P} R_{i\alpha}. \quad (5)$$

The smaller the value of R , the higher the recommendation accuracy.

In reality, online systems only present users with only the top part of the recommendation list. Therefore, we further consider another more practical recommendation accuracy measurement called *precision*, which only takes into account each user’s top- L items in the recommendation list. For each user i , his precision of recommendation is calculated as

$$P_i(L) = \frac{d_i(L)}{L}, \quad (6)$$

where $d_i(L)$ represents the number of user i ’s deleted links contained in the top- L places in the recommendation list. For the whole system, the precision $P(L)$ can be obtained by averaging the individual precisions over all users with at least one link in the probe set.

Predicting what a user likes from the list of best sellers is generally easy in recommendation, while uncovering users’ very personalized preference (i.e. uncovering the unpopular items in the probe set) is much more difficult and important. Therefore, diversity should be considered as another significant aspects for recommender systems besides accuracy. In this letter, we employ two kinds of diversity measurement: *interdiversity* and *novelty*.

The interdiversity mainly consider how users’ recommendation lists are different from each other. Here, we measure it by the Hamming distance. Denoting $C_{ij}(L)$ as

Table 1: The performance of different methods in *Movielens* and *NetfliX* data. The recommendation list length is set as $L = 20$. In MD with NM method, the parameter δ is chosen as -0.05 . In the hybrid method, all the metrics are calculated under the optimal parameter as discussed in ref. [9]. The entries corresponding to the best performance over all methods are emphasized in black.

Network	Method	R	$R_{k \leq 5}$	$P(20)$	$H(20)$	$N(20)$
Movielens	Original MD	0.0933	0.7324	0.1427	0.7161	303.8
	MD with NM	0.0670	0.4089	0.2720	0.8451	254.8
	Hybrid	0.0759	0.5059	0.1532	0.8055	276.7
NetfliX	Original MD	0.0457	0.5618	0.0886	0.5443	2803
	MD with NM	0.0454	0.4713	0.0918	0.5613	2784
	Hybrid	0.0450	0.5174	0.0885	0.5469	2806

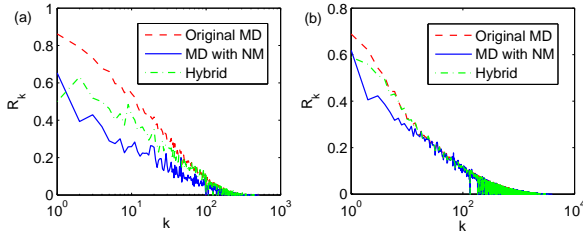


Fig. 3: (Color online) R_k vs. item degree k when using different methods in (a) *Movielens* and (b) *NetfliX* data sets.

the number of common items in the top- L place of the recommendation list of user i and j , their hamming distance can be calculated as

$$H_{ij}(L) = 1 - \frac{C_{ij}(L)}{L}. \quad (7)$$

Clearly, $H_{ij}(L)$ is between 0 and 1, which are respectively corresponding to the cases where i and j having the same or entirely different recommendation lists. Again, averaging $H_{ij}(L)$ over all pairs of users, we obtain the mean hamming distance $H(L)$. A more personalized recommendation results in a higher $H(L)$.

The novelty measures the the average degree of the items in the recommendation list. For those popular items, users may already get them from other channels. However, it's hard for the users to find the relevant but unpopular item. Therefore, a good recommender system should prefer to recommend small degree items. The metric *novelty* can be expressed as

$$N_i(L) = \frac{1}{L} \sum_{\alpha \in O_i} k_\alpha \quad (8)$$

where O_i represents the recommendation list for user i . A low mean popularity $N(L)$ for the whole system indicates a high novel and unexpected recommendation of items.

Results. — We will begin our analysis with the recommendation accuracy since it is one of the most important aspect to evaluate the recommendation results. We

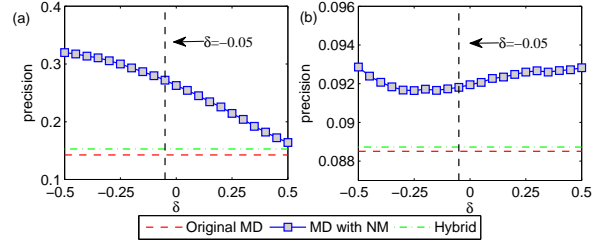


Fig. 4: (Color online) The precision under different δ when $Q = 2$ in (a) *Movielens* and $Q = 1$ in (b) *NetfliX*. The vertical dash line is the optimal δ we used.

first investigate the result of ranking score R under different δ and Q . Since the NM method partially aims at solving the cold-start problem, we also define a local ranking score which is the average ranking score of the items with degree not larger than 5 (denoted as $R_{k \leq 5}$). The results on *Movielens* and *NetfliX* data are reported in Fig. 1. Clearly, with more links added to the network, $R_{k \leq 5}$ becomes smaller. Consequently, the overall R is improved. Given the value of Q , a smaller δ yields a lower $R_{k \leq 5}$, which means the recommendation for the small degree items becomes more accurate. However, the overall R does not monotonously change with δ . For each Q , there is a corresponding optimal δ which yields the best R .

For solving the cold-start problem, fig. 1(c) and (d) suggest that large Q generally works better. However, in order to keep a reasonable computational time of the recommendation algorithm, we select $Q = 2$ in *movielens* data and $Q = 1$ in *NetfliX* data. In the following discussion, we will keep $Q = 2$ and $Q = 1$ in *movielens* and *netfliX* data, respectively. The results of the ranking score are studied more detailedly in fig. 2. In this figure, beside the curve of the MD with NM, we plot the results of the original MD and Hybrid algorithms (without adding any virtual links) as a comparison. For overall R , negative δ clearly works better. However, positive δ is beneficial for improving the ranking score for small degree items. In our simulation, we find δ near 0 ($\delta^* = -0.05$) performs best. As we can see from fig. 2, under this δ^* , the

Table 2: The performance of different methods when the recommendation list length varies in *MovieLens* and *NetfliX* data. In MD with NM method, the parameter δ is chosen as -0.05 . In the hybrid method, all the metrics are calculated under the optimal parameter as discussed in ref. [9]. The entries corresponding to the best performance over all methods are emphasized in black.

Network	Method	$P(50)$	$P(100)$	$H(50)$	$H(100)$	$N(50)$	$N(100)$
MovieLens	Original MD	0.0948	0.0646	0.6395	0.5418	252.1	215.6
	MD with NM	0.1565	0.0913	0.8078	0.7523	203.5	171.9
	Hybrid	0.1031	0.0707	0.7699	0.7299	220.9	178.2
NetfliX	Original MD	0.0594	0.0419	0.4222	0.3496	2337	1876
	MD with NM	0.0621	0.0450	0.4334	0.3784	2326	1860
	Hybrid	0.0596	0.0420	0.4243	0.3522	2335	1874

overall R can remarkably outperform not only the original MD algorithm but also the hybrid algorithm in movieLens data. In Netflix data, δ^* yields a similar R to the original MD and hybrid algorithms. However, the MD with MN method outperforms the other two algorithms in $R_{k \leq 5}$ in both data sets. For the value of each metric, see table I.

To show how the ranking score varies on items with different value of degrees, we additionally investigate an item-degree-dependent ranking score R_k [25]. R_k is defined as the average ranking score over items with the same value of degrees. In fig. 3, the relation between R_k and the item degree k is displayed respectively for the MovieLens and Netflix at the optimal parameters $\delta^* = -0.05$. Besides the MD with NM method, we also plot the results of original MD and hybrid method for comparison. Obviously, the ranking score of small degree items can be significantly improved by adding the virtual connections. Moreover, the ranking score of large degree items can be effectively preserved.

As discuss above, another way to estimate the accuracy of the recommendation results is the precision. Here, we select the recommendation list $L = 20$, and report the precision of MD with NM, original MD and hybrid methods in fig. 4. In ref. [9], it is shown that the hybrid approach can improve the precision compared to the original MD method, so that the green line is higher than the red in fig. 4. In the MD with NM method, a negative δ generally works better for precision in both data sets, which is consistent with the case of the overall R . Furthermore, we observe that the MD with NM method can largely outperform the other two algorithms in precision under the optimal δ^* .

In additional to accuracy, the recommendation diversity is of great significance. For interdiversity, we can estimate how the recommendation results are different from user to user. A larger hamming distance indicates a more personalized recommendation. Besides, the novelty is also an important aspect. With a small novelty, the average degree of the recommended items are low, so that more fresh items will appear in the recommendation list. Setting the recommendation list length as $L = 20$, the related results

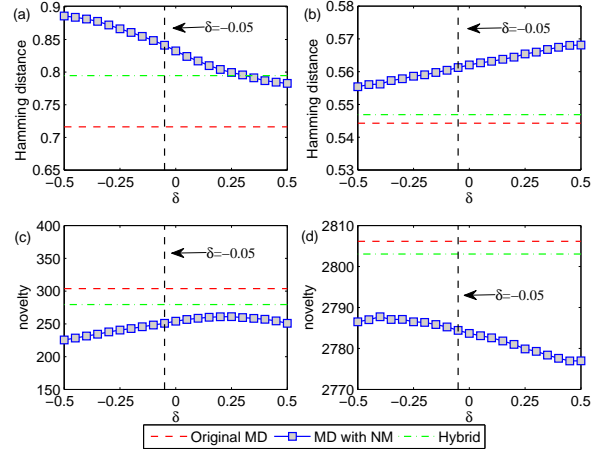


Fig. 5: (Color online) The hamming distance and novelty under different δ when $Q = 2$ in (a), (c) MovieLens and $Q = 1$ in (b), (d) Netflix. The recommendation list length is set as $L = 20$. The vertical dash line is the optimal δ we used.

of different methods are reported in fig. 5 and the detailed value can be seen in table I.

In fig.5, we immediately notice that the MD with NM method yields a bigger hamming distance and smaller novelty index compared to the original MD and hybrid method. These results indicate that our method provide more diverse recommendation results for users. When adding virtual links, the parameter δ actually controls the preference of the virtual links to items with different degree. With a positive δ , more virtual links are added to small degree items and small degree items will be promoted to higher places in users' recommendation list. Accordingly, hamming distance (novelty) in principle should increase (decrease) with δ as shown in Netflix data in Fig. 5. However, we remark that the effect of δ on these two metrics are not always monotonous since these two metrics are focusing only on a small number of top ranked items in the recommendation list. In movieLens data, we can even observe that a small positive δ yields a small hamming distance and a large novelty. However, if we increase δ to

a certain larger positive value in our simulation, the hamming distance will increase and the novelty will decrease in the movielens data.

In reality, the recommendation list length L varies in different online commercial systems. Therefore, we further investigate the cases where $L = 50$ and $L = 100$ in the metrics including precision, hamming distance and novelty. The results are reported in table II. Consistent with the above case where $L = 20$, the MD with NM method outperforms the original MD and hybrid method, which suggests that the improvement from the NM method is very robust.

Conclusion. – Information abundant is a serious problem nowadays for online users. In order to filter irrelevant information, many recommendation algorithms have been proposed. In this field, one of the biggest challenges is the cold-start problem, i.e. the new items have too little historical record to be correctly recommended. So far, all the methods dedicated to solve this problem focused on modifying the existing methods by introducing some parameters. In this letter, we try to solve the problem by directly adding some virtual connection to the bipartite networks so that the niche items have enough information for the recommendation algorithms. Interestingly, besides improving the recommendation accuracy (especially for small degree items), our method can enhance the recommendation diversity compared to the well-known hybrid method of mass diffusion and heat conduction algorithms.

In practice, it is actually not necessary to add too many virtual links to the networks if we only want to enhance the accuracy for those not so popular items and preserve more or less the overall recommendation accuracy. Generally, adding 10% links will be sufficient to solve the cold-start problem (see fig. 1). Therefore, our method can be easily applied to real online commercial systems without increasing too much the computational complexity of the recommendation process. Finally, whether the current NM method is the optimal one for each recommendation algorithm is still unknown. For instance, some special algorithms such as the heat conduction algorithm, which mainly recommends niche items, might require for a different virtual link adding strategy. Related problems ask for further investigation in the future.

We would like to thank Prof. Yi-Cheng Zhang for helpful suggestions. This work is supported by the Foundation of Jiangxi Provincial Department of Education (GJJ. 10696).

REFERENCES

- [1] BRODER A., KUMAR R., MOGHOUL F., RAGHAVAN P., RAJAGOPALAN S., STATA R., TOMKINS A. and WIENER J., *Comput. Netw.*, **33** (2000) 309.
- [2] ADOMAVICIUS G. and TUZHILIN A., *IEEE Trans. Know. Data Eng.*, **17** (2005) 734.
- [3] CACHEDA F., CARNEIRO V., FERNÁNDEZ D. and FORMOSO V., *ACM Trans. Web*, **5** (2011) 1.
- [4] LU L., MEDO M., YEUNG C. H., ZHANG Y.-C., ZHANG Z.-K. and ZHOU T., *Physics Report*, (2012) 10.1016/j.physrep.2012.02.006.
- [5] KONSTAN J. A., MILLER B. N., MALTZ D., HERLOCKER J. L., GORDON L. R. and RIEDL J., *Commun. ACM*, **40** (1997) 77.
- [6] HERLOCKER J. L., KONSTAN J. A., TERVEEN K. and RIEDL J. T., *ACM Trans. Inf. Syst. secur.*, **22** (2004) 5.
- [7] ZHOU T., REN J., MEDO M. and ZHANG Y.-C., *Phys. Rev. E*, **76** (2007) 046115.
- [8] ZHANG Y.-C., BLATTNER M. and YU Y.-K., *Phys. Rev. Lett.*, **99** (2007) 154301.
- [9] ZHOU T., KUSCSIK Z., LIU J.-G., MEDO M., WAKELING J. R. and ZHANG Y.-C., *Proc. Natl. Acad. Sci.*, **107** (2010) 4511.
- [10] ZENG A., YEUNG C. H., SHANG M.-S. and ZHANG Y.-C., *Europhys. Lett.*, **97** (2012) 18005.
- [11] LU L. and LIU W., *Phys. Rev. E*, **83** (2011) 066119.
- [12] QIU T., CHEN G., ZHANG Z.-K. and ZHOU T., *Europhys. Lett.*, **95** (2011) 58003.
- [13] LIU J.-G., ZHOU T. and GUO Q., *Phys. Rev. E*, **84** (2011) 037101.
- [14] EVANS T. S. and PLATO A. D. K., *Phys. Rev. E*, **75** (2007) 056101.
- [15] ZHANG C.-J. and ZENG A., *Physica A*, **391** (2012) 1822.
- [16] NISHIKAWA T. and MOTTER A. E., *Proc. Natl. Acad. Sci.*, **107** (2010) 10342.
- [17] ZENG A., LU L. and ZHOU T., *New J. Phys.*, **14** (2012) 083006.
- [18] YANG H., NIE Y., ZENG A., FAN Y., HU Y. AND DI Z., *Europhys. Lett.*, **89** (2010) 58002.
- [19] SCHNEIDER C. M., MOREIRA A. A., ANDRADE JR. J. S., HAVLIN S. and HERRMANN H. J., *Proc. Natl. Acad. Sci.*, **108** (2011) 3838.
- [20] ZENG A. and LIU W., *Phys. Rev. E*, **85** (2012) 066130.
- [21] LI G., REIS S. D. S., MOREIRA A. A., HAVLIN S., STANLEY H. E. and ANDRADE JR. J. S., *Phys. Rev. Lett.*, **104** (2010) 018701.
- [22] [HTTP://WWW.GROUPLENS.ORG/](http://www.grouplens.org/),
- [23] [HTTP://WWW.NETFLIXPRIZE.COM/](http://www.netflixprize.com/),
- [24] SALTON G. and MCGILL M. J., *MuGraw-Hill, Auckland*, (1983) .
- [25] ZHOU T., JIANG L.-L., SU R.-Q. and ZHANG Y.-C., *Europhys. Lett.*, **81** (2008) 58004.